全球标本数字化建设及共享发展趋势

陈建平[1]，许哲平[2*]

（1.上海辰山植物园，上海 201602；2.中国科学院文献情报中心，北京 100190）

摘要：标本数字化建设是生物多样性保护和利用的重要工作基础，通过标本数据的整合分析，能够在生物分类学、生态学、生物工程、生物保护、粮食安全、生物多样性评估、教学教育和人类社会活动等方面提供数据支撑。为了了解全球标本数字化建设工作的现状以及数据共享的策略与技术发展趋势，通过对比为中国的标本数字化建设工作提供建议，该文分别调查梳理了北美洲、南美洲、欧洲、非洲、亚洲和大洋洲地区的标本数字化和平台建设情况，对标本数据共享现状和趋势从数据使用协议、新技术新方法和公众科学等方面进行了分析，为中国国内的标本数字化工作提出了工作建议，包括：加强标本数字化建设、管理和动态更新方面的协同机制建设，确保实物资源和数字化资源信息的同步；加强数据整理和发布，促进数据质量的提升，充分开放数据使用协议，减少数据使用的阻碍；加强对新技术的学习和引入，特别是开源软件、机器学习和人工智能技术的应用，能够在标签快速识别、自动鉴定和属性数据提取等方面发挥作用；加强区域和国际合作，推动数据的整合应用；推动公众科学项目发展，促进野外采集、室内整理、在线纠错、数据产品研发等工作的开展。

关键字：标本数字化,数据共享,GBIF,公众科学,生物多样性

Global specimen digitization and sharing trends

CHEN Jianping[1]，XU Zheping[2*]

(1. *Chenshan Botanical Garden*, Shanghai 201602, China; 2. *National Science Library, Chinese Academy of Sciences,* Beijing 100190, China)

收稿日期：2021-05-13

基金项目：科技部基础性工作专项（2015FY110200）；中国科学院文献情报领域引进优秀人才计划；中国科学院A类战略性先导科技专项（XDA19050000）［Supported by a grant from the Basic Work Special Project of the National Ministry of Science and Technology of China (2015FY110200); Talent Introducing Plan in the field of Library and Information Science in the Chinese Academy of Sciences; Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA19050000)］。

作者简介：陈建平（1974-），硕士，高级工程师，长期从事生物多样性信息学研究及信息平台建设工作,（E-mail）chenjianping@csnbgsh.cn。

***通讯作者：**许哲平，博士，副研究馆员，长期从事生物多样性信息学以及科学数据管理与利用的研究工作，（E-mail）xuzp@mail.las.ac.cn。

Abstract: The digitization of specimens is an important basis for the conservation and utilization of biodiversity. Through the integrated analysis of specimen data, it can provide data support in taxonomy, ecology, bioengineering, biological protection, food security, biodiversity assessment, human social activities and education and other aspects. At present, the development situation varies from country to country. In order to understand the current status of global specimen digitization work, as well as data sharing strategies and technology development trends, through comparison, provide suggestions for China's specimen digitization work, this article summarizes the status of specimen digitization and platform construction in North America, South America, Europe, Africa, Asia and Oceania, and reviews the status and trends of specimen data sharing from data use agreements, new technologies and methods, and citizen science using. After comparison and analysis with the current situation in China, proposed work suggestions, including strengthening the construction of coordination mechanisms in the digital construction, management and dynamic update of specimens, ensuring the synchronization of physical resources and digital resource information; strengthening data collation and publishing, promoting data quality improvement, fully opening data use agreements, and reducing data use obstacles; strengthen the learning and introduction of new technologies, especially the application of open source software, machine learning and artificial intelligence technologies, which can play a role in rapid tag identification, automatic identification and attribute data extraction; strengthen regional and international cooperation to promote data. The integration and application of data products; promote the development of citizen science projects, and promote the development of field collection, indoor sorting, online error correction, and data product research and development.

Key words: specimen digitization, data sharing, GBIF, citizen science, biodiversity

1. 引言

过去 450 年的时间里，科研人员收集的植物标本数量超过 3.81 亿，分布在全球 3000 多个标本馆中 (Krishtalka et al., 2016; Thiers, 2017)。通过标本数据的整合分析，能够在生物分类学、生态学、生物工程、生物保护、粮食安全、生物多样性评估、人类社会活动和教学教育等方面提供重要支撑(Culley et al.,2013; Heberling et al.,2017; Soltis et al., 2017; Willis et al., 2017;马克平等，2018;张健，2017)。在 2012 年，GBIF 在综合众多专家意见的基础上发布了全球生物多样性信息学展望报告(GBIO)（Hobern et al.,2012）。该报告从文化、数据、实证和知识理解四个层次对未来全球的生物多样性数据相关的研究做了展望，并将标本采集数据作为五类基础数据源之一（其他四类数据源为出版物材料、野外观测、基因测序和自动化遥感观测），为标本数据的整合和应用指明了方向。

在全球范围内，过去十年通过 GBIF（生物多样性信息机构）、CoL（全球生物物种名录）、EOL（网络生命大百科）和 BHL（生物多样性历史文献图书馆）等生物多样性数据项目的推动，生物标本数据的汇聚和共享进展迅速。以全球最大的生物多样性观测数据平台——GBIF

（全球生物多样性信息机构）为例（GBIF，2021），目前观测记录（Occurrences）数据总量为 16.97 亿条，其中传统标本馆的数字化标本数据为 1.85 亿条（占比整个平台数据的 10.9%），其中动物（8637 万条，占比 46.5%）、植物（8586 万条，占比 46.2%）、菌物（707 万条，占比 3.8%）。其中，这些标本数据分布的前 10 个国家是美国（3 546 万条）、巴西（1 258 万条）、澳大利亚（1 216 万条）、墨西哥（892 万条）、加拿大（796 万条）、日本（609 万条）、哥斯达黎加（526 万条）、挪威（452 万条）、西班牙（404 万条）、瑞典（366 万条）。为了进一步加强对全球馆藏资源的整合，GBIF 在 2019 年还启动了全球科学馆藏注册系统项目(GRSciColl)（GBIF,2021），收集研究机构、馆藏和相关工作人员的数据，涵盖了所有相关学科，包括地球与空间科学、人类学、考古学、生物学和生物医学以及农业、兽医学和技术等应用领域，增补了标本数据的元数据与相关背景信息，提升了数据质量。

本文通过梳理世界各大洲的标本数字化建设情况，对标本数据共享现状和趋势进行了调研，与我国当前的建设情况进行对比和分析，在数字化建设与共享服务、协同机制、国际合作和公众科学等方面提出了相应的建议。

2. 数字化建设

2.1 北美洲标本数字化建设

北美的标本数字化以美国的 iDigBio 平台为代表，iDigBio 是一个跨机构合作的综合性生物多样性数据平台，是北美地区标本数字化的门户网站（iDigBio,2021）。目前已经数字化的标本记录数量为 1.28 亿条（植物占 47%），多媒体文件记录 3917 万条（植物占 82.5%），总共 1688 个数据集。2017 年前是 iDigBio 数字化工作开展最为迅猛的阶段。2017 年至今，数据量则呈现出平稳的上升趋势。在项目组织上，iDigBio 将参加单位（数据源）的融入划分为四个阶段，分别是准备、协商、行动、数据汇总，按照参加单位所处的阶段，有序开展工作，使得项目能够稳定持续地推进。

在数据规范与标准上，iDigBio 有详细的标本数字化文件规范、图像存储规范、图像处理规范、图像使用规范，使得数字化工作有明确的操作标准。其规范与标准在 iDigBio 官网上都有开放性的文档说明。基本原则可以概括为：采集图像尽量采用设备的极限分辨率，保证采集质量，图像应该采用无损压缩格式永久存档，图像处理应避免人为过分修饰，处理图像应该基于原始图像，避免误差积累，应尽量为用户提供最佳质量。

在标本数字化技术上，iDigBio 将数字化任务划分为五个核心任务集：

（1）数字化前期：主要是标本实体的修复与规范化整理工作。

（2）图像采集：使用专业单反数码相机或者高清标本扫描仪进行图像采集。

（3）图像处理：包括 9 个常规工作，分别是质量控制、条形码获取、格式转换、颜色亮度调整、图像修剪、图像叠加增强、图像编辑、文件传输、图像内文字识别。

（4）电子数据获取：是指提取或输入标签数据到数据库的过程。具体的工作方法可以是自动化或手工输入，手段包括 OCR 图像文字识别、语音输入、键盘输入等。

（5）地理位置的描述与地标化处理：即将文本描述的地名，审核并标注出精确的经纬度坐标点，建议附上误差范围、坐标系等参数。

2.2 欧洲标本数字化建设

欧洲的植物标本馆历史悠久，馆藏丰富，学术机构众多，机构间也形成了良好的合作关系，其中欧洲分类学联盟 CETAF（Consortium of European Taxonomic Facilities）是最大的分类学研究网络，包括了 5000 多会员，联盟下机构保存了全球 80%已描述的生物多样性标本与数据。BioCASe（The Biological Collection Access Service ） 是 CETAF 推动下创建的数据标准与软件基础体系，指导着标本的数据化与数据共享。BioCASe 数据标准主推 ABCD 标准，在软件系统上提供了全方位的解决方案，包括适用于创建数据平台的 BioCASe Portal 系统，以及适用于数据提供者的 BioCASe Provider Software（BPS）系统，以及网站监控、数据质量检查的工具。BioCASe 是 GBIF 的成员节点，在数据标准、应用系统、共享政策上对 GBIF 有着重要的影响。通过 BioCASe 的数据整合，欧洲各国的标本数据最终通过 GBIF 进行共享发布。除了 CETAF 与 BioCASe 以及 GBIF 外，欧洲各国的标本馆、植物园等分类学相关机构也普遍都创建了自己的网站系统，各类专题性的网络数据库非常丰富，还有大量的数据并未纳入 GBIF 的共享范围。部分数据库因为起步早、积累丰富，已经成为行业内的基础数据库，成为事实上的国际标准与基础平台，如国际植物名称索引（IPNI）等专题数据库。

在俄罗斯，从 2014 年开始，国立莫斯科大学开始进行标本数字化项目（Alexey P. Seregin, 2018），形成了“国家生物系统仓储银行”计划（Seregin A. P. (Ed.), 2021），包括两个子项目，分别是莫斯科数字标本馆和俄罗斯植物分布图集，并通过在 iNaturalist 上创建的“Flora of Russia”项目来维护图库。目前已经累计形成标本 113 万份，图片 111 万张，物种 3.9 万个，地标化数据 66 万条，标签 46 万张、OCR 记录 66 万条。

在法国，国家自然历史博物馆的植物标本馆（馆代码为 P）的数字化建设也积累了丰富的经验（Le Bras et al., 2017）。该标本馆的第一个专业化工具 Vaillant 数据库在 1980 年代中期就开发出来了。从 1993 年开始建设现在的标本数据库 Sonnerat，该数据库目前不仅用来存放博物馆自身馆藏的标本，同时也是一个法语国家的标本馆网络系统（e-ReColNat 项目）。大规模的数字化计划则是从 2008 年资助的 Renobota 项目开始的。下面是该标本馆数字化建设的历史发展表。

表 1　法国国家自然历史博物馆的植物标本馆数字化项目列表

Table 1　List of Herbarium Digital Projects at the National Museum of Natural History of France

项目名称 Project name	时间段（年） Time span (Year)	资助方 Funder	主要结果 Main output
热带亚洲和美洲的莎草科植物数字化项目 Cyperaceae Digitalization Project in Tropical Asia and America	2001~2003	博物馆内部 Internal project	完成 3.1 万份数字化标本。 Completed 31,000 digital specimens
GBIF	2002~2004	GBIF	第一个全球数字化项目，

			超过 5.18 万份标本被数字化，开发了一个模式标本的搜索引擎。 The world's first global digitization project. More than 51,800 specimens have been digitized, and a search engine for type specimens has been developed.
千年种子库项目 Millennium Seed Bank	2004~2008	英国邱园 Kew Garden	完成 3.1 万份数字化标本，并做了精确地地标配准工作。 The digitization of 31,000 specimens with accurate geographic coordinating was completed.
全球植物项目 The Global Plants Initiative	2004~2006：非洲植物倡议（API） 2007~2008：拉丁美洲植物倡议（LAPI） 2009~2015：全球植物倡议（GPI） 2004-2006: African Plants Initiative (API);2007-2008: Latin American Plants Initiative (LAPI)	梅隆基金会 Andrew W. Mellon Foundation.	完成 18.6 万份数字化标本（其中 17.7 万份模式标本） Completed 186,000 specimens digitization (including 177,000 Type specimens)
Lamarck 标本数字化项目 Lamarck Specimen Digitization Project	2004	国家研究中心 The National Centre for Scientific Research	完成 1.9 万份标本数字化 Completed the digitization of 19,000 specimens
Auguste de Saint-	2009	圣保罗植物园研究	9 300 份标本完成数字

Hilaire虚拟植物标本室项目 Auguste de Saint-Hilaire Virtual Herbarium Project		所，环境信息中心，圣保罗市安帕洛佩斯基萨大学基金会 等 Sao Paulo Botanical Garden Research Institute, Environmental Information Center, Amparo Peschisa University Foundation of Sao Paulo, etc.	化，并做了细化处理 Complete the digitization of 9 300 specimens
Renobota 项目 Renobota Project	2008~2013	博物馆内部 Internal project	>500万份标本被数字化，并完成了图像和标签数字化工作。 More than 5 million specimens with images and labels were digitized
Les Herbonautes 项目 Les Herbonautes Project	2012~2019	La Maison de la Chimie基金会，e-ReColNAt，国家自然历史博物馆 La Maison de la Chimie Foundation, e-ReColNAt, National Museum of Natural History	数字化标本 48.8万 Completed 488 000 digitized specimens.
Open Up!	2013	欧盟基金 EU funds	38.5万份被提供到了Europeana平台 385 000 digital specimen records were submitted to the Europeana platform

Reflora 项目 Reflora Project	2014~2016	国家研究中心 The National Centre for Scientific Research	提供 30 万张图片数据，并与里约热内卢联邦大学国家博物馆合作，建立虚拟的植物标本室。 Provide 300 000 picture data, and cooperate with the National Museum of the Federal University of Rio de Janeiro to establish a virtual herbarium.
e-ReColNat	2013-	国家研究中心，国家自然历史博物馆 National Research Center, National Museum of Natural History	数字化植物标本 964 万 Completed the digitization of 9.64 million plant specimens

2.3 非洲标本数字化建设

非洲的植物资源调查与积累起步于殖民地时期宗主国的植物调查，目前有大量非洲标本分散于欧洲各国的标本馆中，随着欧洲标本的数字化工作已经被逐步上网共享，如比利时的皇家中非博物馆，英国邱园的非洲植物计划等，都创建了专题标本数据库，这些数据最终大多通过 GBIF 对外共享。而非洲各国自身的标本馆建设与网络平台还在起步阶段，通过与其他发达国家开展合作研究，很多项目正在展开，标本数字化还有很大的潜力。以南非国家生物多样性研究所(SANBI)为例，已经创建了独立的网站与专题数据库（SANBI,2021），也参加了世界植物在线、千年种子库等国际合作项目，以及非洲植物 POSA 项目、国家植被数据库 NVD 等项目，推动了自有标本资源的数字化，目前公开共享的数据大部分是通过 GBIF 发布。肯尼亚国家博物馆的东非植物标本馆（East African Herbarium）拥有热带非洲最大的植物学收藏，目前拥有 700,000 多条植物标本及相关记录，它是热带非洲最重要的国家级数据中心，研究主要集中在东非植物的分类、分布、开发利用与保护方面。数据管理技术方面，除了 GBIF 平台外，BRAHMS 系统在非洲应用较为广泛，为肯尼亚国家博物馆、南非 BLFU 标本馆等多个机构均提供了技术支持（East African Herbarium, 2021；BLFU, 2021）。

总之，非洲的标本数字化工作基础资料在欧洲，主要通过 GBIF 共享数据，自有标本资料与数字化工作外部依赖性很强，工作空白区较多，未来的潜力很大。

2.4 南美洲标本数字化建设

南美的大量历史标本都保存于美国、欧洲的各个研究所与大学的标本馆中。除了巴西外，各国的标本馆数字化建设程度都比较低，目前能够获取的标本数据基本都来自 GBIF，按植物标本记录统计，巴西约有 700 万，哥伦比亚 139.2 万，秘鲁 80.7 万，阿根廷 72.2 万，玻利维亚 52.6 万。在独立的信息系统建设上，巴西的体系比较完整，很有特色。巴西的标本数字化工作主要体现在 speciesLink 平台中(speciesLink,2021)，截至 2021 年 4 月 15 日已经上线的数据包括了 534 个数据集，1521.9 万在线记录，其中 1129.5 万条记录具有地理坐标，378.5 万记录包含图像，53.2 万模式标本，其中藻类真菌与植物标本合计有 1097 万记录在线。从 2002 年至今，数字化标本量呈现稳定上升的趋势。在数字标本管理系统方面，speciesLink 的精细化实时管理技术十分突出，其指标系统 indicators，可以详细展示每天新增的数据集、标本记录、地标化记录等量化指标。其数据清理系统，可以在线展示出数字标本中存在的问题类型与错误统计，诸如必填字段的空缺、地理坐标缺失、地标错误（如位置在海中）、重复的编号、怀疑出错的学名人名、地名错误等，很多问题都附上了自动化的建议，问题可以逐级展开，可跟踪到具体的标本记录本身，可供随时修订。其工具与应用软件服务十分丰富，提供了十余套专业软件服务，涉及到数据管理、地标化与地图应用、物种数据库管理、物种分布模型、物种查询浏览器插件、网络平台管理、专业量化指标系统等方面。

2.5 大洋洲标本数字化建设

澳大利亚虚拟植物标本馆是一套在线资源库(AVH,2021)，可在线访问澳大利亚和新西兰植物标本数据，总数量超过 666 万份（来自 23 家澳大利亚和新西兰的标本馆）。后随澳大利亚生物图集(ALA)项目的开展，AVH 被合并到 ALA 中一起发展（ALA,2021）。在 ALA 中，与 AVH 同级别的数据合作伙伴还包括在线动物馆藏记录集（OZCAM）、澳大利亚种子银行合作伙伴（ASBP）和默里达令盆地管理局（MDBA）。

新西兰总的植物标本超过 140 万件，拥有世界上最多的南极植物标本数据，约有 64 万件。2011 年，新西兰虚拟标本馆（NZVH）正式启动，这是一个拥有 11 个标本馆数据的虚拟合作网络，可以在线提供 70 万件标本的查询和检索。该系统是由澳大利亚虚拟标本馆（AVH）提供软件和技术支持。随后，该项目也被合并到 AVH 中，并最终成为 ALA 的一部分。

2.6 亚洲地区

截至目前，亚洲地区由于民族多、语言复杂，以及经济和科研工作相对落后的原因，数字化工作还任重道远。尽管中国大陆和台湾、印度、日本、韩国等国在生物多样性数据库建设方面有比较好的基础,，但大多数亚洲国家尚没有完善的生物多样性数据库, 标本数字化建设工作相对落后。以 GBIF 上东南亚国家的标本数量及其贡献国家来看（表 2），该地区的标本绝大多数并不是由本国发布的，而是欧美国家数字化后发布，还有些国家尚无任何数字标本共享，急切需要内外部力量合作，来共同推动该地区的标本数字化建设工作。

表 2 GBIF 上东南亚国家的数字化标本情况（截止 2021 年 04 月 15 日）

Table 2 Digital specimens of Southeast Asian countries on GBIF （Due April 15, 2021)

国家 Country	标本数 Specimens	国家排名	TOP 3 数据发布国家 Top 3 Data publishing country

		Country ranking	国家 1 和数量 No. 1 Country & Specimens	国家 2 和数量 No. 2 Country & Specimens	国家 3 和数量 No. 3 Country & Specimens
泰国 Thailand	483 883	11	美国：193 745 The United States	荷兰：116 601 Netherlands	英国：63 970 United Kingdom
印度尼西亚 Indonesia	1 631 211	9	荷兰：870 899 Netherlands	美国：310 216 The United States	英国：98 720 United Kingdom
柬埔寨 Cambodia	38 257	-	法国：11 232 France	美国：10 450 The United States	日本：5 979 Japan
老挝 Laos	75 349	-	美国：24 701 The United States	英国：19 250 United Kingdom	法国：9 639 France
缅甸 Burma	65 833	-	美国：50 164 The United States	英国：8 043 United Kingdom	荷兰：3 976 Netherlands
马来西亚 Malaysia	425 071	16	荷兰：215 231 Netherlands	美国：149 856 The United States	英国：40 819 United Kingdom
菲律宾 Philippines	558 903	24	美国：401 800 The United States	荷兰：103 426 Netherlands	瑞典：20 949 Sweden
新加坡 Singapore	19 632	-	美国：7 879 The United States	英国：6 071 United Kingdom	荷兰：4 371 Netherlands
东帝汶 Timor-Leste	14 468	-	澳大利亚： Australia 10 598	荷兰：1 945 Netherlands	美国：554 The United States
越南 Vietnam	267 770	7	美国：96 949 The United States	法国：57 283 France	荷兰：21 580 Netherlands

基于此，我国科研人员在亚洲生物多样性保护和数据库网络计划（ABCDNet，2021）的基础上提出了亚洲植物多样性数字化计划（Mapping Asia Plant，简称 MAP），分东南亚、南亚、西亚、中亚、北亚(俄罗斯亚洲部分)和东北亚等 6 个区域推进工作，从文献出版物和标本数据整理入手，正在逐步推进亚洲地区的生物多样性数字化建设与共享合作。

3. 标本数据共享现状和趋势

3.1 GBIF 平台上的数据共享现状

GBIF 是全球标本数据最大的共享平台，通过对 GBIF 上各大洲的参与国家和数据发布情况进行分析统计，可以了解全球标本数据的总体共享情况（表 3）。可以看到欧美国家的数据贡献占了全球的绝大部分，发展中国家的贡献相对有限，与国际经济文化的发展水平密切相关，但发展中国家潜力巨大，将是未来标本数据增长的基础。

表 3　参与 GBIF 数据发布的国家统计

Table 3 Statistics of countries participating in the publishing of GBIF data

区域 Region	国家数 Country number	参与国家列表 Countries	数据发布者 Data providers	数据量（万条） Data volume (10 K)
北美洲 North America	2	美国、加拿大 The United States, Canada	293	56 219.7
欧洲 Europe	21	安道尔、白俄罗斯共和国、比利时、丹麦、爱沙尼亚、芬兰、法国、德国、冰岛、爱尔兰、卢森堡、荷兰、挪威、波兰、葡萄牙、斯洛伐克、斯洛文尼亚、西班牙、瑞典、瑞士、英国 Andorra, Republic of Belarus, Belgium, Denmark, Estonia, Finland, France, Germany, Iceland, Ireland, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom	664	54 726
大洋洲 Oceania	3	澳大利亚、新西兰、汤加 Australia, New Zealand, Tonga	368	8 647.2
南美洲 South America	8	阿根廷、巴西、智利、哥伦比亚、哥斯达黎加、墨西哥、秘鲁、乌拉圭 Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Peru, Uruguay	368	5 648.4
非洲 Africa	22	安哥拉共和国、贝宁、喀麦隆、中非共和国、刚果民主共和国、厄瓜多尔、加纳、几内亚、肯尼亚、利比里亚、马达加斯加、马拉维、马里、毛里塔尼亚、尼日尔、尼尔利亚、南非、南苏丹、坦桑尼亚、多	159	3 005.1

		哥、乌干达、津巴布韦 Angola, Benin, Cameroon, Central Africa, Congo, Ecuador, Ghana, Guinea, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Niger, Nigeria, South Africa, South Sudan, Tanzania, Togo, Uganda, Zimbabwe		
亚洲 Asia	6	日本、韩国、越南、中国（包括台湾）、菲律宾、尼泊尔 Japan, South Korea, Vietnam, China (including Taiwan), Philippines, Nepal	68	1 736.2
总计 Total	62		1 920	129 982.6

3.2 数据使用协议和声明

标本数据在共享和流通之前，需要明确其使用协议，用户才能合法使用和加工。在 2013 年，GBIF 对 1.2 万个数据集中的 4.16 亿条数据记录做了数据协议的总体分析后发现，只有 10%的数据集拥有协议声明。而数据协议竟然有 432 种，这极大地阻碍了数据的共享和流通（Peter Desmet,2013）。在这个调研的基础上，GBIF 管理委员会进行了广泛的沟通和咨询，对混乱的数据协议做了梳理，要求将所有的现有协议都等同地设置为以下三个协议中：CC0、CC BY 和 CC BY-NC。经过梳理之后，目前的数据协议占比情况是：CC0 1.0（56.7%）、CC BY 4.0（27.6%）、CC BY-NC 4.0（15.7%）。通过对北美和澳大利亚的标本数据平台的共享协议（表 4）分析，可以看出 CC0 和 CC BY 是最受欢迎的共享协议。但目前仍然有大量的标本馆平台，数据界面没有明确标识使用协议，部分数据使用还需要繁琐的线下申请和审批流程，制约了数据流通和再利用。

表 4　典型标本平台的数据使用协议

Table 4 Data usage agreement of the main specimen platforms

平台名称 Platform name	国家或区域 Region	数据量 Data volume	共享协议类别 Data usage agreement
GBIF	全球 Global	166×10^6	CC0 1.0、CC BY 4.0、CC BY-NC 4.0
iDigBio	北美 North America	120×10^6	Public-domain or CC0、CC BY、CC BY-SA、CC BY-NC 和 CC BY-NC-SA
AVH	澳大利亚 Australia	0.66×10^6	CC BY 4.0

3.3 新技术和新方法

在标本数字化与数据共享的技术与方法上，受到 IT 技术突飞猛进的影响，在几乎所有应用环节都出现了革新。在数字影像的获取上，流水线式作业的高速扫描系统使得海量标本的快速数字化成为可能，在提升图像质量方面，针对分类学研究的需要，为了强化分类学特征，出现了高清扫描仪、标本的侧光摄影术以及结合解剖镜的显微摄影。在动物标本数字化中，还出现了三维高清影像获取设备等新创举。在野外调查工作中，高清数码相机、智能手机、手持 GPS 等智能设备应用普遍，获取了海量的带有精确地理坐标的植物活体影像数据，为标本提供了丰富的背景资料，某些情况下甚至替代标本成为唯一的凭证资料。

在数据管理与发布方面，除了 GBIF 官方提供的 IPT（集成发布工具包）之外，还有 BRAHMS (Botanical Research and Herbarium Management System，植物研究和标本馆管理系统)应用比较广泛，由英国牛津大学植物系历经数十年研制而成，广泛用于标本馆、植物园、树木园、种子库等科研单位，可提供数据管理与在线发布。

在数据挖掘与分析方面，基于 GBIF 数据已有大量的数据分析工具和代码，可对标本和观测数据进行分析和挖掘，仅 GitHub 网站上，GBIF 相关的开源代码库就有 686 个，其中 R 语言 85 个，Python 语言 54 个，Java 语言 52 个，JavaScript 语言 51 个。 iDigBio、BioCASe、AVH、NSII 等平台也都有专门的应用工具专栏。不同开发语言的数据访问工具包也逐步完善，可以通过数据接口与应用程序快速整合，提供灵活高效的编程环境。例如 iDigBio 提供的 Python 开发包，使得编程访问 iDigBio 数据非常顺畅，加上 Python 科学计算与人工智能的应用生态，可为数据开发者提供强大的技术支持。国内的软件工具开发领域也正在繁荣，如用于标本馆管理的 herblabel（ZHANG et al.，2017），用于分类树构建与分析的 Taxonomic Tree Tool 在线工具（Taxonomic Tree Tool, 2021），用于生物多样性数据清洗、统计与分析的 ipybd (Ipybd,2021)，在技术实力，以及应用效果上都令人印象深刻，具有极强的实用性。

随着人工智能深度学习技术的突破，图像识别 App 进入了实用期，互联网巨头如谷歌、微软、腾讯、百度等均有专题应用，也公布了开放的 API，人工智能识别已经成为一项公共基础服务。应用于生物学领域，就出现了一批生物图像识别的 App，如识别植物的形色、花伴侣，用于识别标本的标本馆伴侣等，虽然在覆盖的物种数量、识别精度上还处于起步阶段，但已经在公众科学、科普领域取得了全社会的关注与认可，对常见植物的识别准确率已经能够满足普通公众的日常需求，通过实践证明了人工智能巨大的应用潜力。未来还需要在识别中加入地理因素，充分利用分类学知识，尤其是标本数据，强化训练，扩大识别范围，让人工智能技术获得更多的应用场景。另外在种子鉴定、花粉鉴定、品种鉴定、有害入侵物种自动鉴定等细分领域也将大有可为，让分类学知识通过人工智能服务于社会。标本数字化提供的大量数据，将是人工智能时代机器学习的重要基础。

3.4 公众科学的发展

标本数字化工作和公众科学探索意愿的结合，催生了各种公众科学项目的诞生。如邱园网站提供了 19 世纪信件识别、真菌特征补全、植物和真菌标本标签识别和手机珍稀植物保护计划等公众可参与项目。大英历史博物馆开展了兰花观察者（Orchid Observers）的公众科学项目，通过公众参与获得数据，使用兰花作为模型，用于深入研究气候对英国植物区系的影响。

北美地区的 iDigBio 平台也提出了公众可参与的众包项目以及 LiveScience 项目。目前影响最大的项目是 iNatrualist，它通过移动 App 为野外考察提供了非常便捷的工具，创建了在线社区，将分类学家与生物爱好者基于兴趣组织起来，通过众包方式收集了大量的生物影像资料，为科学研究提供了丰富的第一手资料。这些项目拉近了公众与标本馆、数据库的距离，将科学研究、科普宣传与社会服务结合在一起，通过灵活的工作策略，获得了普遍的认同。

与国际趋势类似，中国的公众科学活动经过多年的发展，也形成了多种层次多种方式的公众科学信息平台。包括基于论坛、微博、即时通讯工具等通用公众信息交流平台上的兴趣群组，也包括基于移动 App、微信小程序、Web App 等专用工具的兴趣群组。例如中国自然标本馆（CFH）、中国植物图像库（PPBC）、Biotracks、形色、花伴侣、绿途等，均有大量的科学家与爱好者加入。

这些公众科学平台积累了大量的野外观测资料，是对传统标本数据重要的补充，部分资料可以代替标本成为研究的重要凭证。通过整理与审核，其中高质量的数据可以作为标本数据库中的新型数据资源类型，成为标本数字化资源库的正式组成部分。

4. 讨论和建议

通过梳理全球标本数字化的现状和进展，与国际相比，国内标本数字化建设与数据共享已经取得了非常好的成效，并具有自己的优势，但仍然存在一些问题，需要努力解决。包括：标本数字化多是基于数据汇缴的项目制管理，进行集中共享，缺少后期的分布式数据网络节点建设，形成了中心强，节点弱的格局，数据质量的持续更新缺乏相应的机制支持。数据共享方面存在共享协议不规范，标识不明确，尤其是在多语言的互联网环境下，大规模数据使用授权上，仍然需要线下沟通，对数据复用造成困扰。另外，在新技术应用、国际合作以及公众科学方面，虽有良好的基础，但还缺乏亮点项目与应用。通过与国际趋势的对比，结合实际情况，提出以下建议：

（1）加强数字化建设、管理和动态更新方面的协同机制建设，确保实物资源和数字化资源信息的同步。加强标本数据与其他生物多样性数据的融合，成为这个学科的重要资源拼图。

（2）加强数据整理和发布，促进数据质量的提升，特别是最为重要的分类学和时空信息等。充分开放数据使用协议，利用 CC0 或者 CC BY 协议来减少数据使用的阻碍。充分利用 GBIF IPT 工具对数据对外发布，并通过文献引用的跟踪，分析标本数据在不同领域的应用和服务情况。通过数据的共享服务，通过数据使用，获得数据反馈，获得数据更新，提升数据质量。

（3）加强对新技术的学习和引入，特别是机器学习和人工智能技术的应用，能够在标签快速识别分类、标本自动辅助鉴定和特征属性的数据提取等方面发挥作用。加强针对标本数据的开源代码研究。

（4）加强区域和国际合作，加强数据的汇聚和整合。通过 MAP 等区域或国际合作项目来推动跨国家或跨区域的数据建设和共享，带动薄弱国家的标本数字化建设。

（5）加强公众科学项目的合作和推广，让专业人员和大众爱好者参与进来，促进标本数据的野外采集、室内整理、在线纠错、数据产品研发等工作的开展。

参考文献：

ABCDNet: Asia Biodiversity Conservation and Database Network[EB/OL].(2021-04-15) http://www.abcdn.org

Alexey P. Seregin, 2018. The Largest Digital Herbarium in Russia is now available online[J]. TAXON, 67(2):463 - 467

AVH: The Australasian Virtual Herbarium [EB/OL]. (2021-04-15) https://avh.chah.org.au

Culley, T. M. 2013. Why vouchers matter in botanical research[J]. APPL PLANT SCI 1(11): 1300076.

East African Herbarium [EB/OL]. (2021-04-15) http://eaherbarium.museums.or.ke/

Geo Potts Herbarium (BLFU), 2021. [EB/OL].(2021-4-15) https://herbaria.plants.ox.ac.uk/bol/blfu/

Global Biodiversity Information Facility (GBIF), 2021. GBIF Registry of Scientific Collection(GRSciColl)[EB/OL].(2021-04-15) https://www.gbif.org/en/grscicoll

Global Biodiversity Information Facility (GBIF), 2021. Occurrence Search.([EB/OL].(2021-04-15) https://www.gbif.org/occurrence/search

Hobern D, Apostolico A, Arnaud E, 2012. Global Biodiversity Informatics Outlook: Delivering biodiversity knowledge in the information age [EB/OL]. (2021-04-15) https://doi.org/10.15468/6jxa-yb44

Integrated Digitized Biocollections (iDigBio), 2021. [EB/OL]. (2021-04-15) https://www.idigbio.org/

Ipybd-Powerful Data Cleaner For Biodiversity[EB/OL]. (2021-05-13) https://github.com/leisux/ipybd

J. Mason Heberling, Bonnie L. Isaac, 2017. Herbarium specimens as exaptations: New uses for old collections[J]. AMER J BOT,104:1-3.

Krishtalka L, Dalcin E, Ellis S, 2016. Accelerating the discovery of biocollections data. Copenhagen: GBIF Secretariat.[EB/OL](2021-04-15) http://www.gbif.org/resource/83022

Le Bras G, Pignal M, Jeanson ML, 2017. The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Sci Data. 2017 Feb 14;4:170016. doi: 10.1038/sdata.2017.16.[EB/OL] (2021-4-15) https://pubmed.ncbi.nlm.nih.gov/28195585/

MA KP,2017. Mapping Asia Plants: A cyberinfrastructure for plant diversity in Asia[J]. BIODIV SCI, 25(1):1-2 [马克平,2107. 亚洲植物多样性数字化计划[J]. 生物多样性, 25(1): 1-2]

MA KP, ZHU M, JI LQ,2018. Establishing China Infrastructure for Big Biodiversity Data[J]. BULL CAS, 33(8): 838-845. [马克平, 朱敏, 纪力强, 2018. 中国生物多样性大数据平台建设[J].中国科学院院刊, 33(8): 838-845]

Peter Desmet,2013. Analyzing the licenses of all 11,000+ GBIF registered datasets [EB/OL]. (2021-04-15) http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html

Seregin A. P. (Ed.),2021. Moscow Digital Herbarium: Electronic resource. - Moscow State University, Moscow. [EB/OL]. (2021-04-15) https://plant.depo.msu.ru/

Soltis, P. S. 2017. Digitization of herbaria enables novel research[J]. AMER J BOT 104:1-4.

speciesLink [EB/OL]. (2021-04-15) http://www.splink.org.br/

Taxonomic Tree Tool (TTT)[EB/OL]. (2021-04-15) http://ttt.biodinfo.org/

The South African National Biodiversity Institute (SANBI) Website [EB/OL]. (2021-04-15) https://www.sanbi.org/

Thiers, B. 2017. The World's herbaria 2016: A summary report based on data from Index Herbarium. [EB/OL]. (2021-04-15) http://sweetgum.nybg.org/science/ih/

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018. [EB/OL] (2021-04-15) https://doi.org/10.1038/sdata.2016.18

ZHANG J,2017. Biodiversity science and macroecology in the era of big data[J]. BIODIV SCI,25(4), 355-363. [张健,2017.大数据时代的生物多样性科学与宏生态学[J]. 生物多样性,25(4): 355-363.]

ZHANG JL,ZHU HL,LIU JG,2016. Principles behind designing herbarium specimen labels and the R package 'herblabel'[J]. Biodiv Sci,24(12): 1345-1352.[张金龙,朱慧玲,刘金刚,2016. 植物标本标签设计的原则及R程序包herblabel[J].生物多样性, 24(12): 1345-1352.]